基于流批一体的产品质量实时追溯系统研究

谭龙泉 杨宇橙
中车时代电气股份有限公司，中国·湖南 株洲 412000

摘要：在新能源汽车产业爆发的背景下，产品全生命周期质量追溯对产能优化和成本控制至关重要，而传统离线批处理模式因技术架构局限难以满足工业场景动态实时质量管控需求，为此本研究提出基于流批一体的产品质量实时追溯系统，通过云原生流处理技术，构建从采购供应、生产制造到终端服务的全价值链追溯体系。基于分布式集群环境的性能测试表明，实时追溯系统将数据处理时延从离线批处理的 8 小时缩短至 10 秒内，支持秒级异常零件定位、工艺参数实时变更及物流轨迹自动追踪，有效解决传统模式事后追溯难和实时管控弱的核心问题，为企业构建基于数据驱动决策的智能化质量管控体系。

关键词：流批一体；质量追溯；实时管控

Research on Real time Product Quality Traceability System Based on Batch Flow Integration

Tan Longquan, Yang Yucheng
CRRC Times Electric Co., LTD., China Hunan Zhuzhou 412000

Abstract: Under the explosive growth of the new energy vehicle industry, full life-cycle product quality traceability has become critical for production capacity optimization and cost control. However, traditional offline batch processing models fail to meet dynamic real-time quality management requirements in industrial scenarios due to architectural limitations. To address this, this study proposes a stream-batch unified real-time quality traceability system for electric drive components leveraging cloud-native stream processing technology, establishing an end-to-end traceability framework spanning procurement, manufacturing, and after-sales service. Performance benchmarks in distributed cluster environments demonstrate significant improvements: data processing latency is reduced from 8 hours in batch processing to under 10 seconds in real-time operation, enabling sub-second defective part identification, real-time process parameter adjustments, and automated logistics trajectory tracking. This effectively resolves core challenges of post-facto traceability difficulties and weak real-time control in traditional approaches, empowering enterprises to establish intelligent quality management system based on data-driven decision-making.

Keywords: Stream-batch integration; Quality traceability; Real-time control

1 概述

1.1 研究背景

在“双碳”目标驱动下，新能源汽车与工业自动化领域呈现爆发式增长态势，工业电机市场规模持续扩大，2024 年全球电驱系统市场规模已高达 320 亿美元。产品生产过程包含多道核心工序，涉及 100+ 关键参数，其业务与轨道交通本部业务有较大差异，主要体现在以下四方面：（1）需求协同：同主机厂协同，一方面研发项目协同，一方面对接主机厂批量滚动预测和实际订单，快速响应并满足客户需求；（2）生产物流：少品种，大批量生产模式，针对不同的客户需求需采取常规订单、JIT/JIS 方式生产和交付模式，对于一些关键供应商同样也要采取类似模式进行管理；（3）研发设计：与主机厂协同研发、研发和量产界限清晰，产品标准化程度高，超级 BOM 广泛应用；（4）质量与成本：行业法规对质量管理有强制性要求，如追溯、召回等，微利决定必须精细成本管理，产能爬坡时期的成本监控。构建从“研发 – 生产 – 供应 – 营销 – 服务”产品全生命周期质量的追溯研究体系，实现质量问题精准定位和快速响应，对汽车制造业的高质量可持续发展具有重要意义。

1.2 国内外研究现状

1.2.1 流计算框架发展现状

流计算研究始于 20 世纪 90 年代，IBM 研究院启动 System S 项目，基于分布式存储多级流水线并行处理数据，突破传统批处理模式限制，实现从存储后计算到流动中计算的范式跨越。System S 依赖单一节点进行任务调度，单点故障会直接导致集群崩溃。开源系统 Apache Storm[1] 继承其设计理念，引入 ZooKeeper 去中心化协调管理集群，

实现资源的弹性伸缩，并通过多层次容错机制确保流式数据处理的高可靠性。在同时分析历史批量和实时增量数据的场景下，Apache Flink[4] 首次实现流批一体，即一个模型处理无界流数据和有界数据集，支持存算分离架构，解耦计算任务和状态存储，突破本地磁盘瓶颈，支持计算和存储资源独立扩展。

1.2.2 实时数据库研究现状

从技术架构维度出发，将实时数据库划分为内存数据库，时序数据库和流式数据库三类。其中内存数据库基于全内存数据存储实现微秒级读写访问，单节点容量受限且扩容成本高昂；时序数据库以时间戳为索引，采用列式存储和时间分区策略高效压缩数据，支持高吞吐写入，但多维关联查询能力较弱；流式数据库集成流式计算引擎，毫秒级延迟响应，但开发难度较高。

1.3 技术难点

传统离线批处理模式存在两大技术瓶颈：（1）实时处理能力不足，离线批处理依赖定时任务触发数据处理流程，无法满足工业生产中对实时监控异常数据的需求。（2）时序数据管理冗余，需要引入时间戳字段管理时序数据，数据一致性维护成本高，复杂追溯查询响应缓慢。

1.4 研究内容

本研究聚焦产品全生命周期质量追溯的实时性与精准性需求，基于流批一体技术架构，构建从理论建模到工程实践的完整研究体系。将数据处理时延从小时级缩短至秒级，为企业制造构建数据驱动的智能化质量管控体系提供技术底座，实现质量追溯从事后分析到实时管控的跨越升级。

2 基于流批一体的产品质量实时追溯系统

2.1 需求分析

公司产业定位成为国内领先、国际知名的新能源汽车核心部件供应商及系统集成商；聚焦电驱系统，提供全方位系统解决方案，满足不同层次的客户需求。在公司业务形态上，形成以三大板块为主导的产业布局，分别是：（1）电力电子产品，包括板卡、单电机控制器、双电机控制器；（2）电机及部件，包括电机定转子、电机；（3）电驱系统，包括多平台的二合一动力系统、三合一动力系统。面对产能的快速提升和产线自动化水平的提高，产线实现过程中质量问题需要快速定位、定因，以降低质量异常带来的成本和产能的损失。汽车客户要求产品异常须在 48 小时内完成产品整个生产过程的追溯，72 小时内完成定位。

产品质量追溯系统，旨在以产品序列号、批次号和物料编码为标识，打通“研发设计 – 生产制造 – 供应链协同 – 市场营销 – 售后服务”五大业务领域数据壁垒，构建覆盖产品全生命周期的数据网络，实现从原材料到终端用户的双向精准溯源与生产过程动态管控。根据多源业务数据整合，生成具有实际业务价值的关键指标，进而为企业生产计划制定、工艺参数优化等方面提供帮助指导，助力企业构建全价值量质量管理体系。同时支持单体企业管控模式到“总部 + 多基地”分级管控模式，形成统一的质量数据采集、质量过程控制和监控预警平台，实现质量可知、可控、可提高。

2.2 系统设计

为补齐批处理任务的时效性短板，质量追溯系统引入流计算技术实时同步业务数据。首先，该系统需支持高并发接入制造执行系统（Manufacturing Execution System，MES）、仓储管理系统（Warehouse Management System，WMS）、企业资源计划管理（Enterprise Resource Planning，ERP）等异构数据源；其次，数据通过 Kafka 消息队列中转，提供主题分区优化策略；最后，流计算引擎 Flink 读取数据变更捕获（Change Data Capture，CDC）的增量记录，采用时间戳对齐和水位线技术解决数据乱序问题。

2.2.1 多源异构数据采集模块

如图 2.1 所示，MES、ERP 和 WMS 系统的物料源数据，通过序列号或批次号的追溯标识关联零件与产品，产线过站由工业扫描设备扫描记录标识信息，初步筛选过期、不合格零件。当零件具有唯一标识码时，扫描实物和产品条码；对于批次产出的零件，扫描批次和产品条码，并与生产订单号形成三维绑定。数据采集模块从 ERP 拉取物料基础数据，向 MES 反馈扫描结果以驱动工单状态流转，并与 WMS 实时同步物料消耗数量，将不同产线异构设备的扫描记录存储至 Oracle 数据库。

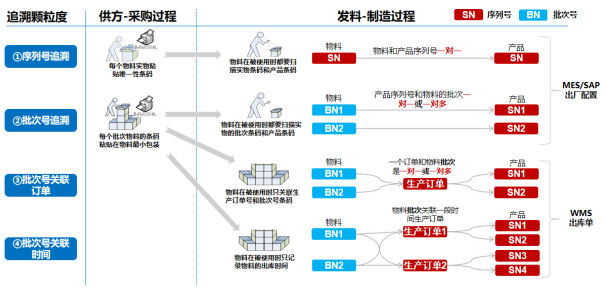

图2.1 数据采集流程图

2.2.2 Kafka 数据传输缓冲模块

本系统基于一表一主题将每个业务实体表映射为独立的 Kafka 主题，为简化管理并确保数据有序性，各主题分区数设置为 1，降低多分区下的消息乱序风险，使用 3 副本策略确保数据存在冗余。生产者产生的数据实时推送到 Kafka 指定主题，存储在 3 个独立 Broker 上，其中 1 个 Broker 作为 Leader 副本处理读写请求，供消费者消费，另

外 2 个作为 Follower 副本实时同步数据，当 Leader 节点发生故障时，其中一个 Follower 节点被选举为新的 Leader，保障服务可用性。

2.2.3 Flink 实时处理分析模块

使用 Flink 实现流批计算一体化，需考虑数据逻辑性、准确性和容错性。根据业务需求使用 Flink DataStream API 对有界批数据和无界流数据统一建模，数据处理逻辑的基本单元被称为算子（Operator），定义数据的输入、转换和输出操作，实现复杂业务逻辑的分层解耦。一个算子可以接收多个输入流，通过合并与连接实现多源数据融合，也可以生成多个输出流，通过分流与侧流输出分发数据。数据流在算子间传递，形成流批一体的数据模型。

Flink 窗口（Window）将无界数据流动态划分为有限的、离散的数据块，每个窗口内执行聚合运算。水位线（Watermark）是插入数据流的一个时间戳标记，触发窗口的结束边界。根据乱序数据的时间戳分配窗口，保证数据处理的准确性。

Flink 通过状态建立起算子间的关联，并由状态后端（State Backend）统一存储管理。状态后端一方面处理算子的状态读写请求，保障流计算的连续性；另一方面通过定期生成检查点，将状态快照持久化存储，提供故障恢复和状态回滚的容错能力。

2.3 系统改进

相比 Flink，RisingWave 将实现细节封装于系统底层，支持快速构建实时分析功能。首先 RisingWave 基于 PostgreSQL 查询语言自动生成分布式算子链，自动优化执行计划。其次根据应用场景匹配合适的窗口类型，有效降低数据处理延迟，减少冗余计算。常用的窗口类型为滚动窗口 (Tumbling Window)、滑动窗口 (Sliding Window) 和会话窗口 (Session Windows) 三类。除此之外，RisingWave 内置检查点机制，定期生成包含计算节点状态和元数据的全局一致性快照，并动态调整检查点间隔和清理过期快照，发生故障可快速定位最新快照并恢复节点状态，保证实时任务的稳定性和可靠性。

3 产品质量实时追溯系统实现

3.1 环境搭建

本文研究基于流批一体的产品质量追溯系统，其硬件环境基于株洲中车电气时代股份有限公司信息中心搭建的分布式集群，当一个节点发生故障时其他节点仍可以提供服务，保证系统持续稳定运行。

3.2 数仓开发

业务系统源数据经贴源层采集、数仓层转换和应用层封装，将离散的业务数据转换为可驱动决策的数据资产。其中数仓层承接贴源层原始数据，完成数据转换、多源关联和聚合计算，为应用层提供结构化明细数据。

3.2.1 贴源层

贴源层（Operational Data Store，ODS）接入 TMES、XMES、WMS 和 ERP 多源异构系统，通过 Kafka 实时增量采集业务源数据，并向 RisingWave 数据库输送。RisingWave 数据库中执行过滤、去重等基础操作，保留原始字段，存储结构化数据。

3.2.2 数仓层

数仓层（Data Warehouse，DW）深度加工贴源层数据，以实体的唯一标识码跨系统关联数据，开发通用业务逻辑模型。首先从贴源层业务系统中提取维度信息，构建维度表；其次根据业务流程设计事实表，形成多表关联的明细数据集；最后基于 RisingWave 数据库搭建流式处理链汇总计算细粒度数据。考虑实时处理链的动态特性，底层数据结构的变动会触发上层视图级联重构，因此模型设计时需要兼顾业务敏捷性与系统稳定性，缩小底层字段变更的影响范围，减少因业务需求变动造成的运维负担。

基于以上考虑，本系统开发数仓层模型时基于状态建立多张物化视图替换单一复杂视图，分类管理零件生产、仓储和运送的不同状态，并为每个状态附加动态属性，实现对零件的全生命周期实时追溯。

3.2.3 应用层

应用层（Application Data Store，ADS）直接面向终端用户，屏蔽数仓层的数据加工细节，生成业务驱动信息。与传统离线数仓相比，实时应用层更强调数据输出的低延迟性，支持用户即时查询最新状态。

3.3 性能测试

本小节与基于传统批处理的产品质量追溯系统对比性能，以计算频率和计算耗时为衡量标准。如表 3.1 所示，传统批处理依赖任务调度，计算全量数据，单次处理周期平均高达 8 小时，业务系统当日产生的数据需要延至次日完成计算，这种滞后性影响企业的生产制造、仓储管理和物流运输，当产线出现问题零件混入、仓库库存数量异常增减或运输途中发生意外时，系统无法基于实时数据快速定位问题零件的生产批次、工艺参数及流向信息，一方面需要人工跨系统核查，耗时费力，另一方面造成生产资源的严重浪费，且影响企业质量管控效率。而基于流批一体的产品质量追溯系统秒级响应输入的每一条数据，计算延迟缩短至 10s 内，仓储库存波动与物流轨迹偏差可实时同步至追溯系统，并对产线异常零件快速锁定装配产品的多

级信息，将质量问题的扩散范围从事后全量筛查压缩至分钟级精准拦截，解决传统模式下“事后追溯难、实时管控弱”的显著痛点。

表3.1 批处理与流批一体性能测试对比

计算测试	时延	
	批处理	流批一体
平均频率	1次/天	1条/次
平均耗时	8h	<10s

3.4 应用效果

3.4.1 业务支撑

在追溯总览界面，查询产品描述、物料描述等基础数据及实物状态和数量动态数据，支持通过交互式操作跳转至明细页面，快速索引出总成零件信息，实现从发现问题零件到关联总成产品和生产批次的秒级追溯。

3.4.2 效益分析

（1）满足主机厂客户需求，在系统中输入零件的序列号、物料编码或箱号任意一项，可即时获取该零件的追溯总览、生产详情、入库溯源和出库追踪信息，以追溯标识实现产品端到端全链路质量管控。

（2）单次追溯时间：当发生客户投诉或出现质量问题时，工程师需要快速追溯有效信息来排查、拦截问题。在大批量制造的背景下，8h/ 次的追溯效率已远远不能满足质量管控要求，极易引起批量性质量风险。系统通过自动化抽取、清洗、关联数据，单次追溯耗时由 8 小时缩短到 5 分钟，效率提升 96 倍。

（3）节约追溯月均耗时：工程师进行问题分析时，80% 的时间用于数据收集，20% 的时间用于数据分析。减少数据收集时间，能大大提升组织的问题分析能效。系统运行 1 年累计追溯 4.1 万次，节约追溯月均耗时 854h/m。

4 结论与展望

4.1 结论

本研究立足新能源产品质量动态管控的迫切需求，提出基于流批一体的产品质量实时追溯系统，打通研发设计、生产制造、供应链协同、市场营销和售后服务五大核心业务领域数据壁垒，构建覆盖产品全生命周期的数据网络。在技术架构方面，借助 Kafka 高并发实时采集 WMS、TMES 和 XMES 等多源异构系统数据，结合 RisingWave 流式数据库的存算分离架构和自动管理机制，统一实时数据流和历史数据批的处理范式，突破传统批处理在实时性的技术瓶颈，将数据处理时延从小时级大幅缩短至秒级。在工业应用方面，支持以产品序列号或批次号为标识的精确追溯与批次追溯双模式，通过追溯总览、生产详情、入库溯源和出库追踪界面，实现对异常生产零件的秒级定位和库存数量变化的实时同步，助力新能源汽车制造的数字化转型，最终达到降本增效的战略目标。

4.2 展望

未来借助 AI 技术摆脱人工干预的现状。当零件出现异常时，尽管实时追溯系统能有效控制影响范围，但在异常发现、原因定位到策略制定的流程中，仍依赖人工介入，导致产线停滞时间较长，资源调配效率较低。通过 AI 技术与追溯系统的深度协同，自动识别异常特征的同时进行风险预警，并提供可解释的原因分析报告，从事后被动响应转变为事前主动预防，实现系统智能化升级。

参考文献：

[1] Iqbal M H, Soomro T R. Big data analysis: Apache storm perspective[J]. International Journal of Computer Trends and Technology, 2015, 19(1): 9-14.

[2] Long H H, Rao R N, Miao W S, et al. An improved topology schedule algorithm for storm system[C]// Computer Science and Applications: Proceedings of the 2014 Asia-Pacific Conference on Computer Science and Applications (CSAC 2014). Shanghai, China, 2014: 187.

[3] Aniello L, Baldoni R, Querzoni L. Adaptive online scheduling in Storm[C]//Proceedings of the 7th ACM International Conference on Distributed Event-Based Systems (DEBS). ACM, 2013: 207-218.

[4] Carbone P, Katsifodimos A, Ewen S, et al. Apache flink: Stream and batch processing in a single engine[J]. The Bulletin of the Technical Committee on Data Engineering, 2015, 38(4).

[5] Chen C, Li K L, et al. FlinkCL: An OpenCL-based in-memory computing architecture on heterogeneous CPU-GPU clusters for big data[J]. IEEE Transactions on Computers, 2018, 67(12): 1765-1779.

[6] Färber F, May N, Lehner W, et al. The SAP HANA Database—An Architecture Overview[J]. IEEE Data Engineering Bulletin, 2012, 35(1): 28-33.

[7] 陈晓宇 , 杨川 , 胡陈啸. 深入浅出 Prometheus：原理、应用、源码与拓展详解 [M]. 北京 : 电子工业出版社 , 2019: 60-79.

[8] Kreps J, Narkhede N, Rao J. Kafka: A distributed messaging system for log processing[C]//Proceedings of the NetDB. 2011, 11(2011): 1-7.

作者简介：谭龙泉（1986.10-），男，汉族，湖南省茶陵县人，大学本科，工程师，数据技术组长，研究方向：数据中台技术。